学　号:200323041　　UDC:_____

厦 门 大 学

硕 士 学 位 论 文

反褶积模型的混合分布估计及其 BOOTSTRAP 置信区间

Nonparametric Estimates of Mixture Distribution and Bootstrap Confidence Interval for Deconvolution Models

吴 绍 凤

指导教师姓名: 林建华　教授

厦门大学数学科学学院

专业名称: 概率论与数理统计

论文提交日期: 2006 年 5 月

论文答辩日期: 2006 年 6 月

学位授予日期: 2005 年　月

答辩委员会主席:_____

评阅人:_____

2006 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的责任。

声明人 (签名):

年　月　日

目 录

Contents

摘　要

在工程技术、生物医学以及其它许多实际领域中，存在一些不可直接观测的变量。因为由于自然环境或者问题本质的限制，这个变量的观测值通常带有误差。如果需要从这个变量的观测数据来推测相关问题的性质时，采用反褶积模型来估计这个变量的分布或密度函数显的尤为重要。本文的做法是将这一未知分布函数设为高斯混合分布的形式，对其中参数进行估计，以解决这一变量的分布的估计问题。本文对这一问题采用 bootstrap 模拟方法得出分布函数的估计，并进一步建立该分布函数的非参 bootstrap 百分位区间。在数值试验中将我们的处理方式与传统的 EM 算法得到的分布估计和正态逼近区间作比较，数值结果表明用 bootstrap 模拟方法得到的准确度更好，数值效果更理想。

关键词：反卷积；混合分布模型；EM 算法；Bootstrap。

Abstract

There are many unobservable variables in some practical fields, For this, deconvolution and mixture distribution has been developed and most widely used. In this paper, we consider the estimation of a distribution function when observations from this distribution are contaminated by measurement error. The approach for using mixture distributions and bootstrap simulations is used to solve this problem, For two parts, distribution function and confidence interval, we show that our result is much better than Clifford, B. C.

Key word: Deconvolution, Mixture Distribution, Expectation-Maximization Algorithm, Bootstrap。

第一章 引言

在实际问题中，常常需要通过获得一个随机变量 X 的观测值来考察它的相关性质，最基本的是通过 X 的观测值推断它所服从的分布函数 $F_X(x)$ 或者相应的密度函数 $f_X(x)$。但在现实中由于测量设备或自然环境等因素的影响，X 的真实值不可能直接获得，所观测随机变量 X 的值通常带有误差。事实上，所测得的值的形式为 $Y = X + \xi$，这里 ξ 是表示测量误差的随机变量。

这类情况大量存在于生物医药、工程技术和经济金融等实际领域。例如在 AIDS 试验中，需要测得某种病毒从感染到症状开始出现所需时间，但实际上得到的数值则是从感染到症状已出现后某一点的时间，而真实的值不可能直接测得。对这一问题的解决，需要用到反褶积模型 [1][2]。

下面给出反褶积的定义。

设 $X_1, \ldots, X_n$ 是一组从未知分布 F_X 中独立抽取实值随机变量，考虑以下形式的观测值

$$Y_i = X_i + \xi_i, \quad i = 1, \ldots, n. \tag{1}$$

这里 $\{\xi_i, i = 1, \ldots, n\}$ 称为误差变量，是服从分布 F_ξ 的独立随机变量，且与 $\{X_i, i = 1, \ldots, n\}$ 相互独立。此时，Y_i 的分布函数可以表示为

$$F_Y(x) = \int_R F_X(x - u) \cdot dF_\xi(u). \tag{2}$$

若 ξ 的密度函数存在，则相应的密度褶积为

$$f_Y(x) = \int_R f_X(x - u) \cdot f_\xi(u) du.$$

反褶积模型要解决的问题之一，也是本文的基本任务: 就是由这组独立观测值 $Y_1, \ldots, Y_n$，以及褶积分布形式 $f_Y(x)$, 来估计随机变量 X 的分布 $f_X(x)$ ，并保证一定的相合性和收敛性。

反褶积的另外一个应用是在非参回归问题中 [3]。设 (X,Z) 是一个随机变量对，考虑回归函数 $m(x)=E(Z|X=x)$ 的估计问题。在实际观测中，应变量 Z 与自变量 X 的值都可能受到误差干扰，比如所观测的值不是所期望的 $(X_1,Z_1),\ldots,(X_n,Z_n)$，而是 $(Y_1,Z_1),\ldots,(Y_n,Z_n)$，其中 Y_i 具有式 (1) 的形式。如何利用这组数据建立一个非参回归函数的估计 $\hat{m}(x)$，使它既具备一般回归函数的性质，又可以体现出变量带有误差，解决这一问题也要用到反褶积的知识。由于此内容不是本文讨论的重点，这里就不在详细论述。

二 研究框架

本文的结构框架如下：

第二章是文献回顾部分，是对下一章将要讨论的内容做铺垫和提供理论基础，可分为三块内容。第一节主要介绍反褶积模型的估计方法如核估计，小波估计等以及它们的收敛性问题，并引出有限混合分布模型。第二节着重介绍了有限混合分布模型的两种处理方法：贝叶斯 (Bayes) 估计和极大似然估计 (MLE) 等，Bayes 部分侧重于后验分布估计以及 MCMC(Markov Chain Monte Carlo) 算法的介绍，而极大似然估计则重点介绍了 EM 算法 (Expectation Maximum Algorithm) 和它的收敛率问题。第三节介绍了 Bootstrap 的发展概况和基本应用，包括标准差的参数和非参估计和几种置信区间的建立等。

第三章是全文的主体，反褶积问题的处理采用的是高斯混合正态分布模型，主要内容是反褶积模型的混合分布估计及其 Bootstrap 百分位置信区间。第一节首先对高斯混合分布模型给出简介，并针对这个模型提出适合的 EM 算法估计混合比例这一参数，进而得到未知分布的估计 $\widehat{F}_X$；然后回顾了 Clifford. B. C. 在文 [31] 中提出的关于 $\widehat{F}_X$ 的正态渐进区间。第二节提出对第一节 EM 算法采用 Bootstrap 进行参数估计，得到更为准确的分布函数的估计 $\overline{F}_X$，并进

一步讨论如何建立 $\widehat{F}_X$ 的 Bootstrap 百分位置信区间。第三节是数值模拟，首先产生一组符合要求的数据，通过比较两种分布函数的估计和两种置信带的模拟结果，证明应用 Bootstrap 的优势，然后用一个实际数据证实了本文提出方法的精确性。最后一节给出了全文的总结，并提出本文方法对解决随机系数回归模型问题的可能性。

本文程序用 S 语言编写，分析图则由 S-Plus 画出。

第二章 文献回顾

第一节 反褶积模型

反褶积模型所要解决的是一类有关测量误差的变量问题，因其广泛的实际背景和适用范围，吸引许多学者注意并加以深入研究。这些研究工作和成果主要集中在两个方面: X 的分布密度估计 $\hat{f}_X(x)$ 以及它的相合性和收敛性。下面简单回顾这两方面的研究成果。

一 核密度估计

对未知分布密度估计最广泛的方法是核密度估计 [4][5]，自然反褶积分布估计研究问题较常用的方法也是核密度估计。它的基本思想是将 $f_X(x)$ 先作核密度估计，然后运用傅立叶 (Fourier) 变换来处理方程 (2)。Carroll Hall(1998)[6]，Fan(1991a)[7] 等对这一估计问题进行了深入研究，下面仅给出 Fan 的估计思想。

(1) 将 X 的分布密度记为

$$f_X(x) = (1/2\pi) \int exp(-itx) \cdot \{\phi_Y(t)/\phi_\xi(t)\}dt,$$

其中 ϕ_Y 与 ϕ_ξ 分别为 Y 与 ξ 的特征函数，ξ 是误差变量。

(2) 由于 Y 分布未知，通常用它的经验特征函数 $\hat{\phi}_n(t) = 1/n \sum_{k=0}^{n} \exp(itY_k)$ 或它的核估计密度 $\hat{\phi}_Y(y) = (1/nh) \sum_{k=0}^{n} K\{(y - Y_k)/h\}$ 来代替，这里 $K(\cdot)$ 为核函数。

(3) 对 (2) 中核函数 $K(\cdot)$ 做傅立叶变换，得到 $\phi_K(\cdot)$，即 $\phi_K(t) = \int exp(itx) \cdot K(x)dx$。

(4) 变量 X 的分布密度核估计就可以写为:

$$\hat{f}_X(x,h) = (1/nh) \sum_{k=0}^{n} g((x - Y_j)/h, h),$$

其中

$$g(x,h)=(1/2\pi)\int exp(-itx)\cdot\{\phi_K(t)/\phi_\xi(t/h)\}dt,$$

h 是核函数 $K(\cdot)$ 的参数，称为窗宽或带宽。

同时 fan 得出结论，如果步骤 (3) 中傅立叶变换的核函数所选带宽范围是有限的，也就是核函数的支撑是有界的，则 X 的核密度估计函数 $\hat{f}_X(x,h)$ 在点和域上的收敛率都可以达到渐进最优。

二 收敛性问题

影响非参反褶积分布密度估计收敛速率 [9][10] 的因素很多，其中误差变量 ξ 的分布是最一个重要的影响因素，如 Fan 在文 [7] 中研究了两种不同的误差分布下的收敛率。他指出，随机误差变量 ξ 分布的平滑度影响了这个估计的收敛性以及非参反褶积估计的难度。

根据平滑度，可以将分布分为两类：一般平滑和超平滑，它们定义如下

称随机变量 η 具有 β 阶超平滑分布，如果它的特征函数 $\phi_\eta(t)$ 满足；

$$d_0|t|^{-\beta_0}exp(-|t|^{-\beta}/\tau)\le|\phi_\eta(t)|\le d_1|t|^{-\beta_1}exp(-|t|^{-\beta}/\tau),$$

其中 d_0,d_1,β,τ 为正的常数，β_0,β_1 为常数；

称随机变量 η 具有 β 阶一般平滑分布，如果它的特征函数 $\phi_\eta(t)$ 满足；

$$d_0|t|^{-\beta}\le|\phi_\eta(t)|\le d_1|t|^{-\beta},$$

其中 d_0,d_1,β 为正的常数。

根据定义，在常见的分布中，正态分布、混合正态分布和柯西分布等属于超平滑分布，伽玛分布、双指数分布以及对称伽玛分布则是一般平滑分布。如超平滑分布

$$\begin{cases} N(0,1) & \beta=2,\\ \pi^{-1}(1+x^2)^{-1}(Cauchy(0,1)) & \beta=1. \end{cases}$$

一般平滑分布

$$\begin{cases} \alpha^p/\Gamma(p)x^{p-1}e^{\alpha x}(gamma) & \beta = p, \\ 2^{-1}e^{-|x|}(double\ \ exponential) & \beta = 2. \end{cases}$$

文 [7] 认为根据随机误差分布两种不同的平滑程度，也存在两种不同类型的最优收敛率。但一般平滑性越高，收敛速度越缓慢，非参反褶积分布估计问题越困难。

三 其它方法

1 小波 (wavelet) 估计

小波展开式也是估计未知分布的一种重要方法。简单介绍它的定义 (更详细的定义参见文献 [8],[9])：设 $\varphi(\cdot) \in L^2(\mathbb{R})$ 与 $\psi(\cdot) \in L^2(\mathbb{R})$ 表示对应于多分辨分析假定下的尺度函数和母小波函数，则 $\{\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k);\ k = 0, \pm 1, \ldots\}$ 与 $\{\psi_{l,k}(x) = 2^{l/2}\psi(2^l x - k);\ l \geq j,\ k = 0, \pm 1, \ldots\}$ (任意 $j = 0, \pm 1, \ldots$) 构成了 $L^2(\mathbb{R})$ 的一组标准正交基。即任意函数 $f(x) \in L^2(\mathbb{R})$ 都可以分解为

$$f(x) = \sum_{k \in Z} a_{j,k}\varphi_{j,k}(x) + \sum_{k \in Z}\sum_{l=k}^{\infty} b_{l,k}\psi_{l,k}(x), \tag{3}$$

这里系数 $a_{j,k}, b_{l,k}$ 分别为 $a_{j,k} = \int_{-\infty}^{\infty} \varphi_{j,k}(x) \cdot f(x)dx$, $b_{l,k} = \int_{-\infty}^{\infty} \psi_{l,k}(x) \cdot\ f(x)dx$.

Marianna.p[10] 等应用它来估计非参反褶积密度。他的基本思想是将待估函数 $f_X(t)$ 写为 Meyer-Wavelet 展开式的形式，然后由反褶积算法来估计展式的系数，从而得到 X 的 wavelet 分布密度估计 $f_{\tilde{X}}(t)$。同时证明了 wavelet 估计对误差变量超平滑分布在 $MISE(t) = E\int(\tilde{f} - f)^2 dt \longrightarrow 0$ 的意义下非参反褶积密度估计有较高的收敛率。

2 混合分布模型

Maritz 和 Lwin(1989)[11] 提出可以用有限混合分布模型估计非参反褶积分布问题。混合分布模型自上个世纪初被提出后，由于它有着广泛的实际应用背景而备受关注，相关的理论研究也较为健全。所谓混合分布，是将分布函数 $F(x)$ 记为

$$F(x)=\sum_{k=1}^{m}\omega_k\cdot\ F_k(x).$$

其中，$F(u)$ 称为成分分布；$\{\omega_k,k\ =\ 1,\ldots,m\}$ 称为混合比例，满足 $\omega_k\ \geq 0,\sum_{k=1}^{m}\omega_k=1;m\in Z$。

混合分布模型与反褶积分布估计问题结合起来，使得二者都有了更广阔的发展空间 [2][12]。本文将对此作进一步讨论，在下一节中将对混合分布模型的发展概况和已存在的经典结论进行简单回顾和总结。

第二节 有限混合分布模型

一 模型概述

有限混合分布模型 [13] 在近 20 年来越来越受到统计学家的重视。首先它不仅为有关总体差异质性如何建立模型提供了一个自然的研究框架，并建立其与聚类判别分析等之间的紧密联系。更重要的，它的出现去除了束缚在未知分布形状上的种种限制，为那些不能由任何单个参数分布族逼近的未知分布估计问题提供了一种异常灵活有效的途径。因而，它成为分布估计的一种重要方法。

为以下叙述方便，给出有限混合分布模型的基本定义。$x_1,\ldots,x_n$ 为一组独立同分布的随机变量，设它们的概率密度函数形式为

$$f(x)=\sum_{k=1}^{m}\omega_k f_k(x). \tag{4}$$

这一形式就称为有限混合分布模型。其中 $\{\omega_k,k=1,\ldots,m\}$ 是混合比例，满足 $\omega_k\geq 0,\sum_{k=1}^{m}\omega_k=1$。

f_k 是混合成分分布对应的密度函数，它一般分为两种情况：(1). f_k 是已知的分布；(2). f_k 形式已知，但其中参数未定，即已知 f_k 属于某个特定分布族，但分布组参数 θ_k 随 k 不同而变化。这种情况下，f_k 常记为 $f_k(x|\theta_k)$，例如我们常见的两个成分的对数正态混合分布可以写为：

$$f(x) = \omega f_1(x) + (1-\omega) f_2(x), \quad 0 \le \omega \le 1.$$

其中

$$f_i(x) = (\sqrt{2\pi}\beta_i x)^{-1} \exp[-(\ln x - \alpha_i)^2 / 2\beta_i^2], i = 1, 2.$$

这里 α_i 和 β_i 分别为第 i 个成分分布中 $\ln x$ 的期望和标准差。

通常，在 m 未知的情况下，混合分布模型的估计问题多采用 bayes 相关的理论和 MCMC(Markov Chain Monte Carlo) 方法解决；在固定的 m 值下，人们也常用极大似然估计 (MLE) 的相关理论以及以此为基础的 EM(Expectation-Maximization) 算法来处理。

二 Bayes 理论的应用

自从 1994 年 Diebolt 和 Robert 在文 [14] 中对模型 (4) 采用 Gibbs 抽样算法在 k 已知条件下估计参数 $\Theta = (\omega_k, \theta_k; \ k = 1, \ldots, m)$ 以来，Bayes 方法在处理混合分布模型问题上有着显著的发展。而接下来的工作大多集中在混合分布的成分个数 m 的研究，这其中包括两个重要的方向，一是采用假设检验 $H_0: \ m \ \mathcal{VS} \ H_1: \ m+1$ 来推断 m 的最优取值；另外一种途径是通过给 m 和参数 Θ 赋予合适的先验分布，通过边际似然等过程得到 m 的后验分布，从而总结它的不确定性 [15][16]。

m 的推断问题说到底就是如何在一堆有竞争力的模型中选择最合适的一个，而应用 Bayes 方法处理此问题，其优点不仅是可以挑选出一个最优的模型，它同时还是综合不同模型结果的一致途径。以至于 Peter Green. 等在文 [17] 中

宣布 Bayes 方法处理混合分布模型问题是非常适合与有效的，尤其是对成分的个数未确定的情况。在 Bayes 框架下处理混合模型的估计问题尽管理论上相对简单，它的计算推导却是相当复杂的。但所幸 MCMC[17][18] 算法的不断创新和发展，和 "Reversible Jump "[18][19] 等概念的提出，以及它们在 Bayes 分析中广泛而日益成熟的应用，解决了这一难题，并且使得混合分布模型的 Bayes 分析向着更深更完善的层次发展。

下面简单介绍有关 m 的后验分析的基本思想，详细内容和推导过程参见文献 [15] 等。

在 Bayes 分析中，当模型 (4) 中 f_k 未知时，三组未知的量 k, ω, θ 均被看成是适当的分布中抽取的，即给它们各自赋予赋予先验分布。通常，$\omega = \{\omega_k, k = 1, \ldots, n\}$ 被认为是服从 Dirichelet 分布，即 $\omega \sim D(\delta_1, \ldots, \delta_m)$；$\theta_k$ 的先验分布为 $\theta_k \sim \Phi_k(\phi)(\phi$ 为参数，分布 $\Phi_k(\phi)$ 的形式根据具体情况选择，比如在混合正态分布中，通常所选的先验分布为 $\alpha_k \sim N(\varepsilon, \kappa^{-1}), \sigma_k^{-2} \sim \Gamma(\alpha, \beta))$；而 m 的先验分布形式记为 $\Pi(k)$。

这样

$$\begin{aligned} & f(x, \theta, g, \omega, m) \\ & = \Pi(k)\Pi(\omega, \delta) f(g|m, \omega)\Pi(\theta|m, \phi) f(x|m, g, \theta). \end{aligned} \tag{5}$$

其中，$g = (g_1, \ldots, g_n)$ 定义如下：g_i 的取值表示第 i 个样本 x_i 是从第几个成分分布中产生的，例如，若第 5 个样本 x_5 是由第 3 个分布生成，则 $g_5 = 3$。

最后经过推导运算，就可以得到 m 的后验分布 $\Pi(m|x)$ 的形式。

三 极大似然估计 (MLE) 与 EM 算法

在模型 (4) 中，确定 m 的值后，混合分布问题就转化为参数 $\Theta = (\omega_k, \theta_k, k =$

$1,\ldots,m$) 的估计问题。许多方法被相继提出或运用到此问题的解决上来，这其中就包括著名的极大似然估计理论。尤其 1960 年以来计算机技术的普及和高速发展，使得极大似然估计对混合分布参数估计问题的重要性越来越显著。

1 极大似然

首先来回顾极大似然估计的定义。

设 $p(x|\Theta)$ 为一密度函数，Θ 为它的参数，而 $\chi=(x_1,\ldots,x_N)$ 是从这个分布中随机抽取的样本，即它们是独立同分布于分布密度 p 的。因此 χ 的密度就可以表示为

$$p(\chi|\Theta)=\prod_{i=1}^{N}p(x_i|\Theta)=L(\Theta|\chi).$$

函数 $L(\Theta|\chi)$ 就称为基于数据 χ 的参数似然，简称似然函数。注意这里 χ 是给定的样本，因而 $L(\Theta|\chi)$ 是参数 Θ 的函数。极大似然估计所要解决的问题是找出一个 Θ 的值使 $L(\Theta|\chi)$ 达到极大。通常为了推导方便，将 $L(\cdot)$ 做对数变换，得到

$$\ell(\Theta|\chi)=\log(L(\Theta|\chi))=\sum_{i=1}^{N}\log p(x_i|\Theta). \tag{6}$$

问题也就转化为求出 Θ^* 满足 $\Theta^*=\arg\max_\Theta L(\Theta|\chi)=\arg\max_\Theta \ell(\Theta|\chi)$。极大似然估计问题的难易程度根据 $p(x|\Theta)$ 的具体形式而定，比如简单的情况，如果 $p(x|\Theta)$ 是一元正态分布密度函数，这时要估计的参数为 $\Theta=(\mu,\sigma)$，就可以通过令 $\ell(\Theta|\chi)$ 的偏导数为 0 来直接求解 Θ 的极大似然估计估计。然而对于更多的问题，却不可能获得 $\ell(\cdot)$ 或其偏导数的分析表达式，这就要求借助于更多复杂而精密的技巧。

2 EM 算法

EM 算法就是这样一个复杂而精密的技巧。它由 ALD(1977)[20] 首次提出，而后又得以不断改进和完善。EM 算法是针对不完全或有缺失数据所属分布中参数的极大似然估计的一种广泛而有效的处理方法。它主要应用在以下两个